Longitudinal Data Analysis with Latent Growth Modeling:

An Introduction and Illustration for Higher Education Researchers

Presented at the American Educational Research Association Annual Conference,

April 8, 2011, New Orleans, LA

Rebecca D. Blanchard, PhD, Baystate Medical Center and Tufts University

Timothy R. Konold, PhD, University of Virginia

Abstract

This paper introduces latent growth modeling (LGM) as a statistical method for analyzing change over time in latent, or unobserved, variables, with particular emphasis of the application of this method in higher education research. While increasingly popular in other areas of education research and despite a wealth of publicly-available datasets relevant to postsecondary education research, LGM has not been utilized widely by higher education researchers. This paper begins by introducing LGM as a desirable mechanism for analyzing variability in individual growth trajectories over time and then presents an illustration of its application. An example of the application of LGM to data obtained from the Integrated Postsecondary Educational Data System (IPEDS) is presented to introduce specific components of LGM, including model specification and goodness-of-fit indices, and to demonstrate the research potential for higher education researchers. Finally, additional datasets offering longitudinal analysis potential for higher education researchers are presented to facilitate research.

Background

Latent growth modeling (LGM) has grown in popularity among educational researchers over the past decade (Marsh & Hau, 2007). LGM relies on a structural equation framework to estimate the growth trajectory for an entire group and the variation within that group, as well as the effectiveness of covariates or predictors to sufficiently explain variation in individual growth trajectories. Due to its flexibility for being applied in various situations and to answer a variety of questions, the methodology has been incorporated into many contexts, including policy analysis (Heck & Takahashi, 2006) and student achievement and development (Konold & Pianta, 2007). However, despite the methodological benefits afforded by LGM, it remains under-utilized in postsecondary education research. The primary goals of this paper are to 1) introduce LGM as a tool for investigating longitudinal data, 2) lead researchers through the analysis and interpretation of a substantive example, and 3) introduce readers to several existing national longitudinal data sets. This paper should be a guide for analyzing data with LGM to address many issues in postsecondary education research.

Illustrative Example

To demonstrate the application of LGM, this paper conducts an illustrative longitudinal investigation of degree rates at a set of colleges and universities using data from the Integrated Postsecondary Education Data System (IPEDS). This illustrative model is aimed at capturing the undergraduate degrees produced in the social sciences and understanding the influence of faculty and student resources on that degree production between 1997 and 2007[1]. The research questions for this example are:

1) To what extent have social science degrees per FTE grown between 1997 and 2007?

2) To what extent do institutions vary in that growth?

[1] This substantive context is for instructive purposes and should not be extrapolated as empirical argument.

3) To what extent does an institution's instruction expenditures and yield rate influence its change in degree rates over the time period?

The sample in this study includes all public and private bachelor's degree-granting institutions classified as doctoral universities, master's colleges or universities, or baccalaureate (arts & sciences) colleges by the Carnegie Classification system[2]. Data collected on these N=1,145 institutions include yield rate, faculty salaries, and awarded degrees (see Table 1). Where appropriate, variables are normalized for inclusion in the model by the full time equivalent enrollment (FTE) calculation, widely used by IPEDS[3].

[2] The 2005 Carnegie Classification system is used, consistent with the classification system available in IPEDS.

[3] FTE - Total full time equivalent enrollment is equal to the sum of both undergraduate and graduate (if applicable) FTE. FTE is calculated as the total number of instructional credit hours divided by the average annual credits per degree-seeking student, as defined by IPEDS. For institutions with a semester, trimester, continuous enrollment, or 4-1-4 plan, the undergraduate denominator is 30 and the graduate denominator is 24. For institutions with a quarter plan, the undergraduate denominator is 45 and the graduate denominator is 36.

Table 1. Variables included in the model

Variable (Year)	Definition	Mean	Std. Deviation
Faculty Salaries per FTE (1997)	Sum of institution's salaries and wages paid to employees – faculty, staff, part time, full time, regular employees, and student employees – that conduct instruction. Amount is for the academic year 1997-1998, divided by the full-time equivalent enrollment at the institution	2,483.30	1,915.97
Faculty Salaries per FTE (2002)	Sum of institution's salaries and wages paid to employees – faculty, staff, part time, full time, regular employees, and student employees – that conduct instruction. Amount is for the academic year 2002-2003, divided by the full-time equivalent enrollment at the institution	3,122.18	4,224.65
Faculty Salaries per FTE (2007)	Sum of institution's salaries and wages paid to employees – faculty, staff, part time, full time, regular employees, and student employees – that conduct instruction. Amount is for the academic year 2007-2008, divided by the full-time equivalent enrollment at the institution	3,513.14	3,046.04
Yield Rate (2003)	Ratio of enrolled students to students accepted into the institution during the academic year 2003-2004. Students included in this yield rate entered the institution in fall 2004.	0.4266	0.1635
Degrees per FTE (1997)	Number of baccalaureate degrees awarded in the social sciences during the academic year 1997-1998, divided by the full-time equivalent enrollment at the institution.	.1605	.0628
Degrees per FTE (2002)	Number of baccalaureate degrees awarded in the social sciences during the academic year 2002-2003, divided by the full-time equivalent enrollment at the institution.	.1865	.1047
Degrees per FTE (2007)	Number of baccalaureate degrees awarded in the social sciences during the academic year 2007-2008, divided by the full-time equivalent enrollment at the institution.	.1745	.0745

NOTE: Variables and definitions obtained from IPEDS

Model Specification

Analysis of the illustrative example was carried through two stages; 1) an unconditional model estimating the change in the outcome variable from 1997-2007, and 2) a conditional model estimating the influence of two covariates on the outcome variable. Analysis is conducted with AMOS[4]. Full information maximum likelihood (FIML) estimation is used to obtain parameter estimates and accommodate missing data[5]. In addition, fit of the illustrative models in this paper are evaluated through commonly accepted fit indices provided by major structural equation software packages (Duncan, Duncan, & Strycker, 2006). These include; chi-square, the Tucker-

[4] AMOS is an acronym for Analysis of Moment Structures

[5] FIML defines the population parameters in the model such that they reflect as accurately as possible the mean and covariance matrix of the sample of institutions (Bollen & Curran, 2006).

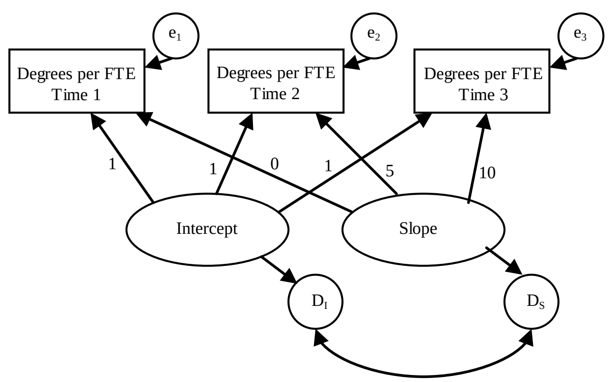
Lewis Index (TLI), the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and Akaike's (1974) information criterion (AIC).

The Unconditional Model

The first step of LGM requires an unconditional model to define growth of the outcome variable over the ten-year time frame (see Figure 1). Consistent with AMOS specification, observed variables are indicated by boxes and latent variables are designated with circles or ovals. Single-headed arrows indicate direct relationships and double-headed arrows (such as that seen between the intercept and slope) represent correlations to be estimated from the data.

The unconditional model provides information about the trend of the outcome variable, including the intercept (i.e., starting point) and slope (i.e., growth) parameters. In addition, LGM allows researchers to capture the variance associated with the growth parameters. These variances, or residual values, (labeled "D_I" and "D_S" in Figure 1) demonstrate how much institutions vary around the group's estimated model parameters.

Figure 1. Unconditional model of undergraduate degrees per FTE from 1997-2007



The Conditional Model

For the second stage of analysis, the unconditional model is expanded to include the two covariate variables hypothesized to influence the output of undergraduate degrees (see Figure 2). Both yield rate and faculty salaries vary across time, so they are included in the model as *time-*

varying covariates (TVCs). In the present model, undergraduate degrees receive effects from these resources and the pattern of growth at each time point[6], as demonstrated by the factor loadings. In addition, the residuals for TVCs between each year are correlated, as these values are likely related over time.

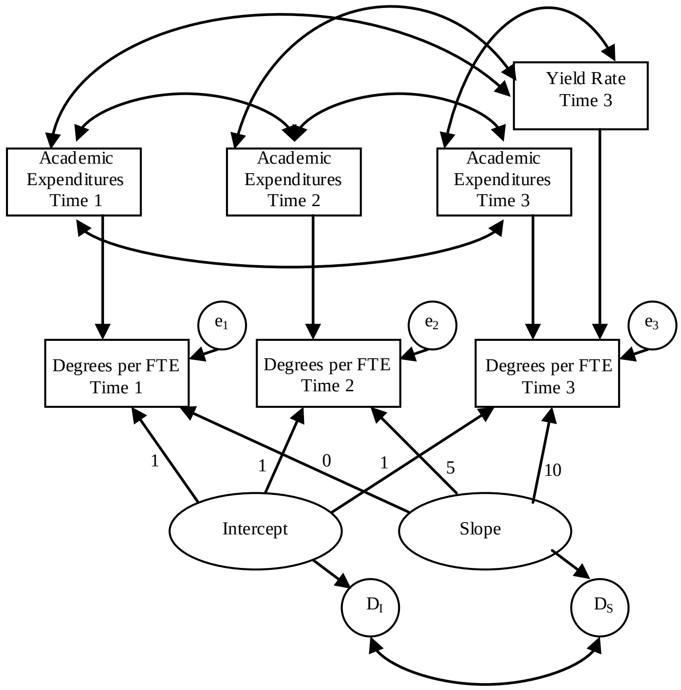
All TVCs are examined with a nested model structure. First, the model is estimated with all TVCs included in the model. Next, the factor loading of one covariate is estimated while the loading for the other is set to zero. Finally, factor loadings for both covariates are estimated. This series of model estimates allows a chi-square test to determine the significance with which each of these iterations improves the fit of the model to the sample data. The most parsimonious model resulting from this series of tests is retained.

The conditional model addresses how TVCs influence the outcome variable of degree rate. This influence is examined with two analysis questions; 1) whether the inclusion of TVCs explains any of the variance (D_I and D_S) in the average slopes or intercepts for undergraduate degree rates and 2) how the inputs are directly related to the output.

---

[6] Because yield rate data was not available for the first two time points, this variable is only added as a covariate in the third time point, attesting to the flexibility of LGM to handle data limitations.

Figure 2. Conditional latent growth model of undergraduate degrees from 1997-2007 with time-varying predictors



Results

Unconditional Model

Results of the analysis are presented in Table 2. Fit statistics were positive, suggesting that the models had reasonable fit to the data. The unconditional model of undergraduate degree rates per FTE answered the first two questions of this study, which asked the extent to which degrees grew over the ten-year period and the degree of variation in that growth.

To answer the first question, the pattern of growth (the linearity or nonlinearity of the growth curve) was tested by constraining and then freely estimating the slope parameters, or pattern coefficients. By constraining the coefficients to 5 and 10 for both of the outputs, the model tested a linear growth between time points. The fit of this model was compared to a non-linear, or *spline*[7],

---

[7] Spline method allows non-linear growth to be modeled more parsimoniously than polynomial growth because fewer parameters are estimated (Bollen & Curran, 2006).

growth model, estimated by anchoring the first and last time points, and allowing the middle time point to be freely estimated (Bollen & Curran, 2006). In Figure 1, the pattern coefficients for undergraduate degrees were set to 0 for 1997 and 10 for 2007, but the slope factor loading for year 2002 was freely estimated from the data. Results indicated that the spline model fit the data better than the linear model.

Table 2. Standardized parameter estimates for unconditional and conditional models

	Unconditional Model	**Conditional Model**
Pattern Coefficients[1]		
1997 (Time 1)	1, 0	1, 0
2002 (Time 2)	1, 32.6 [S]	1, 33.3 [S]
2007 (Time 3)	1, 10	1, 10
Intercept	.172*	.152*
Slope	.001*	.001*
Intercept Variance	.002*	.002*
Slope Variance	.000	.000
Correlation I,S	.481	.392
Time-varying Covariates		
FS on Degrees (Time 1)	-	.114*
FS on Degrees (Time 2)	-	.207*
FS on Degrees (Time 3)	-	.091*
YR on Degrees 3	-	.042*
Fit Statistics		
Chi-sq (df)	17.3 (2)	93.1 (10)
TLI	.899	.935
CFI	.966	.977
RMSEA	.106	.085

NOTES: [1]Coefficients listed as intercept, slope growth or spline ([S]) estimate; *p<.000

As shown in Table 2, the estimated pattern coefficient at time 2 was 32.6 (linear growth would have reflected a parameter at time 2 equal to 5). The rate of growth for the group was significant at .001 degrees per FTE.

Therefore, average growth for undergraduate degrees for this group of institutions over the ten-year period was .010 degrees per FTE (10*.001=.010). From 1997 to 2002, undergraduate growth was .033 degrees per FTE (32.6*.001=.0326). Growth from 2002 to 2007 for undergraduate degrees was (10-32.6)*.001=-.0226. In other words, the number of undergraduate degrees produced decreased from 2002 to 2007 by .023 degrees per FTE. Most of the growth in undergraduate degrees produced per FTE occurred in the first five years of the selected time period (.0326/.010=3.26). Growth then decreased in the last five years (-.023/.010=-2.3). This resulted in the overall increase of 1% in undergraduate degrees per FTE between 1997 and 2007.

The second research question asked the extent to which these institutions varied around the average slope parameter. Though the variation estimate was statistically significant, it was negligible (σ^2_{Slope} <.001), meaning that institutions grew in a similar way over the time period.

Conditional Model

The final question asked to what extent an institution's student and faculty resources influenced its change in degree rates over time. This question was answered with results from the estimated conditional model. As shown in Table 2, inclusion of TVCs did not account for any of the variation around the fixed effects parameters of the model, but they did influence the model in other ways. First, inclusion of the two TVCs slightly improved the model fit, as demonstrated by a chi-square difference test between the full estimated model and the model with the effects of the covariates on the outcome variables set equal to zero (χ^2_D (4)=95.4, p<.001).

Standardized regression weights, or direct effect estimates, for both covariates on the outcome variable were significant across all time points (see Table 2). Controlling for these covariates decreased the average starting point for the full sample of institutions to .152 degrees per FTE but did not change the statistically significant estimated growth. Further, the correlation between the intercept and slope value in this model was slightly less than the unconditional model, but both demonstrated that institutions with high starting values of degrees per FTE increased at a higher rate over the ten-year period.

Next, the direct relationships between the covariates and the outcome variable were analyzed. The coefficients for TVCs are conceptualized as "the time-specific prediction of the repeated measure after controlling for the influence of the underlying growth process," (Bollen & Curran, 2006, p. 194). In other words, the effect of each TVC is interpreted as the influence on the production of undergraduate degrees above and beyond what would be expected as the normal growth in the production of degrees captured by the model. The direct effect influences of faculty salaries and yield rate were significant and positive on undergraduate degrees per FTE across all three time points. This could suggest that, all other factors remaining consistent, an increase in yield rate or faculty salaries could have a small but positive influence on the number of social science degrees produced per FTE.

Longitudinal Datasets

The illustrative example in this paper is one demonstration of the many opportunities to utilize longitudinal data available to higher education researchers. IPEDS is a federally maintained database to which all postsecondary institutions receiving federal aid must report[8]. In addition to institution-level data, panel data is also collected at the student level. For example, national sample surveys conducted by the National Center for Education Statistics (NCES) and the National Science Foundation (NSF) collect data from high school students and their families and teachers, college students, and graduate students. Descriptions of the Education Longitudinal Study (ELS), the Baccalaureate and Beyond (B&B), and the Survey of Earned Doctorates (SED) will be presented in this paper. While access to some of this data is restricted to licensed users, education researchers and graduate students affiliated with institutions should have little trouble gaining access, and would benefit from taking time to get acquainted with the data housed on the NCES and NSF websites. Access to panel data provides a wealth of opportunities for researchers, including easier investigation into patterns of growth and change over time.

[8] The Higher Education Act of 1992 declared that reporting to IPEDS is mandatory for all institutions who participate in federal student financial assistance programs (NCES, n.d.), and currently, over 3,000 public and private higher education institutions report annual data to IPEDs.

Conclusion

This paper demonstrates that LGM offers incredible potential as a tool for longitudinal analysis. First, the method allows examination of average change over time as well as individual variation in that change. Second, covariates can be added to account for variability among the units and then tested for statistical significance. Finally, the prevalence of user-friendly programs such as the SPSS module AMOS makes LGM more accessible to education researchers.

Longitudinal analysis is helpful for higher education researchers to investigate trends in their data which often hold the keys for understanding progress in higher education. This paper walks the researcher through the development of the latent growth model, illustrates an institution-level application of the method, and interprets the results. The researcher is then introduced to a set of national longitudinal datasets which offer many opportunities to explore trends, change, and growth in higher education.

References

Bollen, K.A. & Curran, P.J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Duncan, T.E., Duncan, S.C., & Strycker, L.A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Heck, R.H. & Takahashi, R. (2006). Examining the impact of Proposition 48 on graduation rates in Division 1A football and program recruiting behavior: Testing a policy change model. *Educational Policy*, 20, 587-614.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Konold, T.K. & Pianta, R. C. (2005). Empirically-derived, person-oriented patterns of school readiness in typically developing children: Description and prediction to first grade achievement. *Applied Developmental Science*, 9, 174-187.

Marsh, H.W. & Hau, K.T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology*, 32, 151-170.